R E P O R T R E S U M E S

ED 016 491 LI 000 018

CITATION INDEX AND MEASURES OF ASSOCIATION IN MECHANIZED DOCUMENT RETRIEVAL. RAPPORT NUMMER 2. (REPORT NUMBER 2).

REPORT NUMBER KVAL-PM-295 PUB DATE 25 JAN 67

EDRS PRICE MF-$0.25 HC-$0.64 14P.

DESCRIPTORS- *CITATION INDEXES, *INFORMATION RETRIEVAL, INDEXES (LOCATORS), ALGORITHMS, INFORMATION SYSTEMS, *STATISTICAL ASSOCIATION METHODS,

CITATION INDEXES, IN COMBINATION WITH REGULAR INDEXES, ARE CREATING NEW METHODS OF INFORMATION RETRIEVAL, WHICH USE SOME MEASURE OF CONNECTION OR ASSOCIATION BETWEEN A GIVEN BASE OF ARTICLES AND OTHER DOCUMENTS IN A COLLECTION. GENERAL REQUIREMENTS OF SUCH A MEASURE ARE--(1) IT SHOULD BE INDIRECTLY PROPORTIONAL TO THE NUMBER OF LINKS IN A CHAIN OF CITATIONS BETWEEN THE TWO ARTICLES BEING MEASURED, AND (2) IT SHOULD BE DIRECTLY PROPORTIONAL TO THE NUMBER OF CHAINS CONNECTING THE TWO ARTICLES. THE FIELD OF THERMODYNAMICS OFFERS A SIMPLE PHYSICAL ANALOGY TO THIS MEASURE. MATHEMATICAL PROCEDURES BASED UPON THERMODYNAMICS ARE USED TO DERIVE AN ALGORITHM FOR FINDING THE DEGREE OF ASSOCIATION. (CM)

LI 000 018

ED016491

Språkförlaget
SKRIPTOR
Fack
Stockholm 40
SWEDEN

KVAL PM 295
25.1.1967

AL 000 175

CITATION INDEX AND MEASURES OF ASSOCIATION IN MECHANIZED DOCUMENT RETRIEVAL

Rapport nr 2 till Kungl. Statskontoret

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

LI 000 018

CITATION INDEX AND MEASURES OF ASSOCIATION IN MECHANIZED DOCUMENT RETRIEVAL

1. General Discussion

Everyone with some experience of research work knows the value of the bibliographical references that customarily appear in periodicals and books. These aids are especially important to a scholar who seeks orientation in a field which is new to him.

Normally, it is comparatively easy to find some few articles about the field of interest, and it is easier still to find a book which gives a survey of the field. By aid of the references found in these articles or books, it is usually possible to retrieve other articles, where a more detailed treatment is given of the subject. Now again, in these articles one may get further hints from the bibliographical references contained in them, and so forth. This kind of successive searching - which is, in fact, the recognized retrieval procedure in practical research work - does yield the necessary basic literature for a research project in a surprisingly large number of cases.

The reason for this procedure to be so successful is that there are generally relevant connections between an article and the literature it refers to. After all, the author himself mostly draws up the list of references for his article, i.e., the author states what articles he found most useful for the purpose, and it is reasonable to expect the author to be a better judge than anybody else as to what literature is most closely connected with his own article.

The procedure, however, has one obvious drawback: the references indicate only literature older than the article under inspection. Only exceptionally do authors refer to articles not yet published. The research worker may, however, be more anxious to learn about the recent development of the field than in obtaining a historical survey. That is why "inverse reference lists" - or "citation indexes" as they are often called - have recently been suggested, i.e., lists where for each article are tabulated all later articles referring to it. Such inverse lists, naturally, must be cumulative.

Inverse reference lists are now being published with the support of the U.S. National Science Foundation. One may note how great an importance the "Weinberg Committee"[1] attaches to these experiments. In the summary of its recommendations the committee states:

> The panel wishes to call the attention of the technical community to a promising new method of access to the literature called the citation index: a cumulative list of articles that, subsequent to the appearence of the original article, refer to that article.

No doubt, citation indexes will become important tools for manual information retrieval, especially if used in combination with the customary reference lists. Thus, the scholar may even by zig-zag manœuvres find his way from an article to literature that appeared simultaneously with the first article.

It is now natural to try to design a similar method also for mechanized information retrieval. We shall here suggest such a mechanized procedure where, starting with a number of known articles - the so called "base" - a set of articles is retrieved which, on the evidence of the mutual references between articles, can be expected to be closely connected with the base; of course, we suppose that all references given in the articles in the library have been stored in a computer. Given a set of articles - the "base", supplied by the scholar or a librarian - our retrieval method is supposed to measure the "connections" or "association" between each article in the library and the articles in the base. The articles which then get the highest scores are then accepted as retrieval suggestions. (Naturally, out of these the scholar or the librarian may select a set of "close hits" to serve as a new base in a subsequent computer run.)

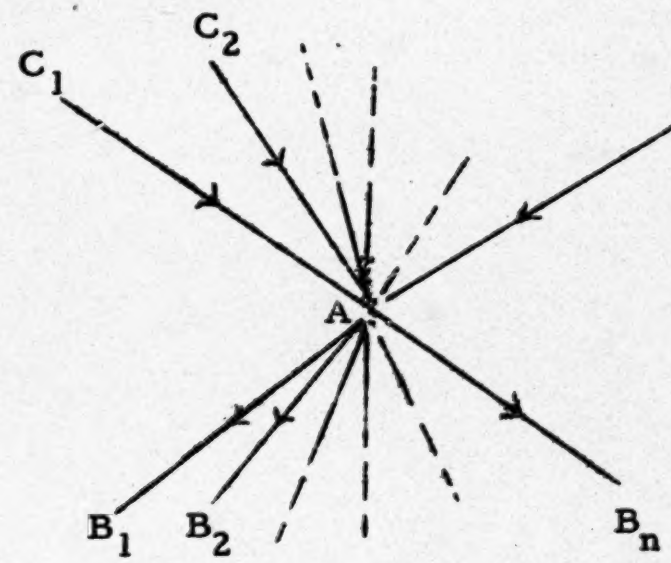
In section 2 below we shall state some very general requirements how such a "measure of associations" should behave, and in section 3 we suggest a specific measure which fulfils those general conditions. This measure is then mathematically treated in section 4.

---

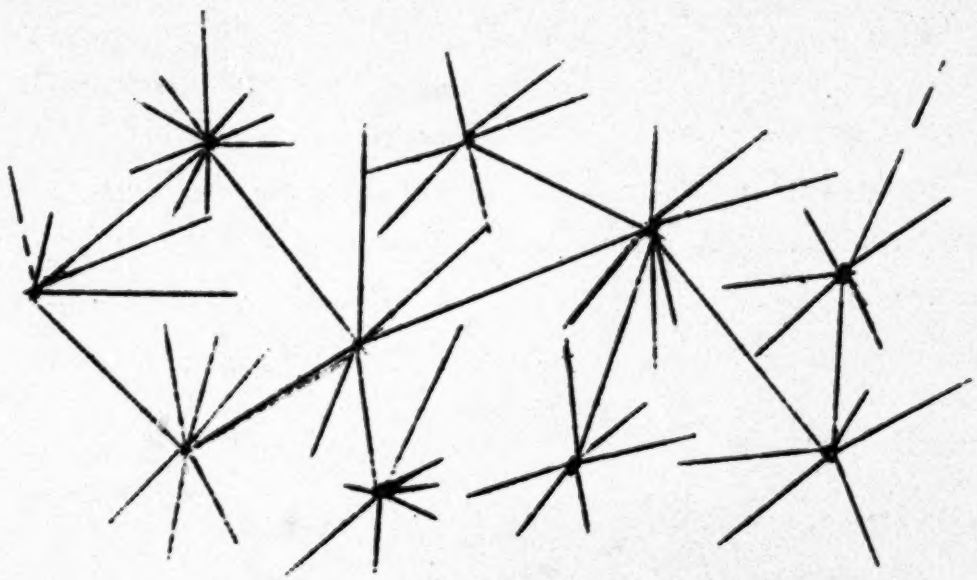
1) The President's Science Advisory Committee, "Science, Government and Information", The White House, January 1963.

2. General Requirements of Measures of Association between Articles

In order to discuss adequate measures of the connection between articles, let us first summarize the relevant features of the library. Every article generally contains some references to articles within the same field and perhaps also some to articles in neighbouring fields. If the article A refers to the article B, we may say that there is an arrow from A pointing at B. Let us suppose that the article A points at the articles $B_1, B_2, \ldots, B_n$. These, then, are the articles mentioned in the reference list of A. Further let us assume that the articles $C_1, C_2, \ldots, C_m$ each point at A. They are, thus, the ones we would find in a citation index under A. We might imagine A as a spider in its web:



Naturally, the articles B_i and C_j, in their turn, are spiders in their webs. The set of articles and the arrows between them will form an utterly complicated topological network:



where the articles are the nodes and the connecting links are the arrows (the references). We may, of course, imagine a time axis in the figure so that articles published later are plotted lower down in the diagram than those published earlier.

We now say that two articles are (directly) connected if there is an arrow from A pointing at B or vice versa. We say that they are indirectly connected if there is a chain of articles $C_1, C_2, \ldots, C_n$, so that A is connected with C_1, C_1 is connected with C_2 and so on until finally C_n is connected with B.

The general idea behind our suggestion is now that it should be possible to define a numerical measure which indicates a degree of connection (or degree of association) between articles so that, e.g., directly connected articles are considered to be more closely associated than indirectly connected articles and so on. To be more specific, we require that a measure should fulfil the following two conditions:

1. Articles connected by a chain consisting of many links should, ceteris paribus, be weaker associated with each other than those connected by a chain of few steps.
2. Articles connected by many chains should, ceteris paribus, be stronger associated than those connected by few chains.

Optionally, the measure should depend on the directions of the arrows in the chains.

Given such a numerical measure - in section 3 and 4 below we shall define one - the following seems to be an effective retrieval procedure.

The scholar arbitrarily assigns numbers to a set of articles (the "base"), these numbers being intended to indicate the actual degree of relevance for him. For each article in the library its measure of association to the base is computed. The articles with the highest association measures are presented to the scholar as being those (probably) most relevant.[1)]

---

1) It is interesting to note that the structure in a library as we have described it is quite similar to a certain semantic structure among words. If we interprete the nodes in our network not as articles but as words and the relation "A points at B" not as "A refers to B" but "A is synonymous to B", we can in a completely analogous way discuss "measures of association" between words. This semantic model can be applied to problems connecting with indexing and definition and construction of thesaurus classes.

3. A Measure of Association

One may now ask if there exists any simple method to define such a measure of associations which fulfils the general conditions 1 and 2 of section 2. The answer is that there do exist several such methods, and we give here what seems to be the simplest one. The proposed measure is one among many others; we have chosen to work out this specific model thoroughly because it yields a numerically simple, programmable algorithm for finding the degree of association between the articles, and because it has a very clear physical analogue, which makes it possible also for a person without a mathematical background to follow the discussion. Thus, the proposed procedure can be made clearer by the aid of the following physical analogy of the structure of articles and arrows.

Let us imagine that the articles are solid particles of some kind, connected by, say, iron rods, thus forming a structure not unlike a modern iron sculpture. Let us assume that there are heat sources, constantly yielding heat to some of the solid particles, the "base". This heat is distributed according to normal physical principles: thus, e.g., heat transmitted from one particle to another will be lost to the first. We also assume that there are constant losses of heat through radiation to the environment. Consequently, an equilibrium will be reached after some time, when the heat losses total up to the same amount as the heat produced by the heat sources. When this equilibrium has been arrived at, the sculpture will, on an average, be cooler farther away from the base. When the system has reached equilibrium, the hottest nodes represent the articles most closely connected with the base, and an enumeration of these will form the output of this information retrieval system.

The physical model just described, however, is not quite adequate for the present purpose. In the model actually used - cf. below, section 4 - the "rods" between the articles behave like semi-conductors, which generally conduct heat better in one direction than in the other. In addition - and this is essential - we introduce a possibility of influencing the permeability constants, that is to say the heat conducting capacity of the arrows. Thus, the searcher may choose a high permeability constant in the forward direction and a low permeability in the backward direction.

This will mean that the retrieval will be concentrated on later literature. In the opposite case, the list of articles recommended will have a character of historical survey. Again, one may choose both constants small but prescribe very high losses by radiation. The computer then suggests articles within a narrow field close to the basic set. Finally, if we put the permeability constant high in both directions and prescribe small amount of radiation, the computer will associate rather freely from article to article.

This possibility of influencing the computer's way of associating by the change of certain parameters can give the information retrieval the form of a game between the scholar and the computer. He can give the computer more or less free reins till he finds something which seems to be interesting. He may then change the parameters so that the computer thereafter proceeds more restrictively.

Naturally, to stimulate such a game between man and machine, it is necessary to arrange for faster feed-back than in to-day's computer systems. The optimal, of course, is to let man and machine work on line with each other via an optical display. This is technically feasible to-day, though discouragingly expensive.

We now turn to the mathematical formulation of an algorithm for computation of association measures and a discussion of the conditions for its application.

4. Mathematical Treatment

We begin by assuming that we have a "library", consisting of a (usually great) number of articles, A_1, A_2, A_3, ...; the enumeration is as far as we are concerned arbitrary. For each of these articles there are certain other articles (usually comparatively few in number) which the first one is "pointing at", i.e., these articles are found in the article's reference list. If the article A_i refers to the article A_j we denote this by "$A_i \rightarrow A_j$".

We now introduce the concept of "time", t, a parameter which we assume to have its range in the non-negative real numbers, and for each i a real-valued function of t, $a_i(t)$, called the "temperature" of the article A_i at the time t. We also assume that for each i = 1, 2, ... we have a fix non-negative number b_i, called the regulating temperature for the article A_i; for convenience, we use a terminology borrowed from the field of thermodynamics, which is likely to be well-known to many readers. Those - comparatively few - articles for which the regulating temperature is different from zero constitute the base.

Now we introduce the following hypotheses on the heat transfer between the articles: The amount of heat transferred between two articles connected by an arrow during a (short) period of time is proportional to the difference between the temperatures of the two articles and to the termal conductivity of the "arrow" in the direction of the flow. The heat energy is always transported from an article with higher temperature to one with lower temperature. [1] At the same time each article has a "gain" of heat energy which is proportional to the difference between its regulating temperature and its actual temperature. In the normal case all temperatures considered are positive. Thus, outside the base the "gain" is in fact a loss, as the regulating temperatures for those articles are all zero.

With these hypotheses it is now quite easy to give the exact form of an equation which governs the heat flow in our library. If we know the temperature, $a_i(t)$, for an article at an instant t we can compute the approximate temperature at the time $(t + \Delta t)$. Thus,

$$(1) \qquad a_i(t+\Delta t) = a_i(t) + \Delta t \{\mu [b_i - a_i(t)]\} + \\ + \Delta t \cdot \alpha \{ \sum_j [a_j(t) - a_i(t)]^- + \sum_k [a_k(t) - a_i(t)]^+ \} + \\ + \Delta t \cdot \beta \{ \sum [a_j(t) - a_i(t)]^+ + \sum [a_k(t) - a_i(t)]^- \},$$

---

1) One may (as we, in fact, have done, though not here) also assume that there is a kind of energy amplification at each node, i.e., the energy-loss at a node is less than the total amount of energy transferred to the neighbours.

$i = 1, 2, \ldots, N$, where in each of the last two lines of the formula the first sum is taken over those $\underline{j}$ for which "$A_i \rightarrow A_j$", and the last sum is taken over those $\underline{k}$ for which "$A_k \rightarrow A_i$". In the formula we have used the denotations "a^+" and "a^-" for:

$$a^+ = \begin{cases} a \text{ if } a \geq 0 \\ 0 \text{ if } a < 0 \end{cases} \qquad \text{and} \qquad a^- = \begin{cases} a \text{ if } a < 0 \\ 0 \text{ if } a \geq 0 \end{cases}$$

With the aid of formula (1) we are able to compute the temperatures for all the articles in the library at any instant when we know the heat distribution a short time before. Thus starting with a given heat distribution at the time $\underline{t} = 0$ (when,we may assume, all the articles have the temperature zero), we may recursively compute the temperatures at closely separated times $t_1, t_2, \ldots, t_l$ and so on. The problem now is whether the temperature stabilizes when $\underline{t}$ converges to infinity, and if so, (Theoreme 3, below), we are interested in what this equilibrium distribution may be (eq. (5) below).

In this paper we do not aim at resolving the intricate questions about the existence and the uniqueness of an equilibrium distribution of the quasi-linear difference scheme (1). However, one can get quite a good view on the general behaviour of a final state distribution by investigating in greater detail the more simple case where $\alpha = \beta$. In that case (1) reduces to an ordinary linear difference equation, which can be treated by more standard manipulations. We state without proof that the results in this paper are to a large extent valid also for the most general case of (1).

Thus, assuming that $\alpha = \beta$ in (1), this equation reduces to

$$(2) \qquad a_i(t + \Delta t) = a_i(t) + \Delta t \{\mu [b_i - a_i(t)]\} + \Delta t\, \alpha \{ \sum_j [a_j(t) - a_i(t)] \}; \quad i = 1, 2, \ldots, N.$$

The summation is now made over all those $\underline{j}$, such that $A_i \rightarrow A_j$ or $A_j \rightarrow A_i$.

In order to study the behaviour of the solution of the recursive equation (2) we introduce some more efficient notations which make the equation (2) more easily handled.

Let A denote the matrix with as many rows and columns as we have articles in the library and with its elements δ_{ik} equal to 1 for those i and k for which "$A_i \rightarrow A_k$" and with all other elements equal to zero. We denote the transpose of A by A^*. (With a mathematical terminology we may say that the matrix A represents the "graph" of the information in the list of references corresponding to the articles of our library and that A^* represents the graph of the information in the citation index.) We introduce two other matrices, B and C, which both are diagonal. B has its i:th element in the diagonal equal to the number of arrows from the article A_i, and C has the corresponding element equal to the number of arrows pointing towards A_i. Finally we define the following two vectors, a(t) and b as

$$a(t) = \begin{bmatrix} a_1(t) \\ a_2(t) \\ a_3(t) \\ \cdot \\ \cdot \\ \cdot \end{bmatrix} \qquad \text{and } b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}$$

With these notations the equation (2) can be rewritten as

$$a(t + \Delta t) = a(t) + \Delta t \{ \mu b + (\alpha A' - \mu E) a(t) \}, \tag{3}$$

where the matrix A′ is equal to $(A + A^* - B - C)$.

In the equation (3) we may shift the first term on the right hand over to the left hand side, divide by Δt and then let Δt tend to zero. The difference equation (3) is then transformed into the following differential equation (with a vector function a(t) as unknown):

$$\frac{d\,a(t)}{dt} = \mu b + (\alpha A' - \mu E) a(t) \tag{4}$$

(E denotes the unit matrix).

The solution of (4) is the vector function $a(t) = (\mu E - \alpha A')^{-1} b +$ $+ \exp\{(\alpha A' - \mu E)t\} \cdot c$. (The vector c is depending on what value

we give the function a(t) for t = 0. For instance, if a(0) = 0, we have $c = (\alpha A' - \mu E)^{-1} \cdot b.$) As is shown below the eigenvalues of the matrix $(\alpha A' - \mu E)$ are all negative when $\mu > 0$, whence the solution of (4) converges (as we wanted it to do) exponentially to a steady state solution.

(5) $$a(\infty) = (\mu E - \alpha A')^{-1} \cdot \mu b.$$

This is the expected equilibrium distribution of temperature in our library, and the articles corresponding to the largest element of a (∞) are those which are to be presented as retrieval suggestions.

In computer applications it is of course very unpractical to define the equilibrium distribution explicitly as in (5), since the matrix to be inversed in this formula is of a very large order (in a more ambitious information retrieval project the order even may exceed one million). Instead we note that a(∞) also may be defined as the solution $\underline{x}$ of the (linear) equation

(6) $$(\mu E - \alpha A')\, x = \mu b$$

The problem is now to obtain an approximate solution of this equation. Before proceeding, though, we want to show that the solution of equation (4) really converges to steady state solution and that this solution also is given by (6). This follows immedately from Theorem 1 below. We also show that this equilibrium distribution behaves as we expect from a real heat distribution. (Theorem 2).

Theorem 1. The eigenvalues of the matrix A′ are all non-positive (A′ is the matrix defined in equation (3) above).

Proof. Let σ be an eigenvalue for the matrix A′ and let $\underline{x}$ be a corresponding eigenvector, i.e., $\underline{x}$ is a solution for the equation $A'x = \sigma x$. In coordinate form this equation may be written as

(7) $$\sum_j (x_j - x_i) = \sigma x_i, \qquad i = 1, 2, \ldots, N$$

For each $\underline{i}$ the sum is extended over j's in a subset of the numbers 1, 2, ..., N. Now assume that $\sigma > 0$ and that $x_i = \max_k \{x_k\}$. If $x_i > 0$, then the right hand side of (7) is > 0 but the left hand side is ≤ 0 (because all terms are non-positive). Thus $\max \{x_k\} \leq 0$. In the same way it follows that $\min \{x_k\} \geq 0$ if σ is assumed to be > 0. But it is impossible that $\max \{x_k\} \leq 0 \leq \min \{x_k\}$ if not all $x_k = 0$. Thus any eigenvalue of the matrix A' is non-positive.

Theorem 2. Assume that $b = (b_1, b_2, \ldots, b_N)^* \neq 0$ is a vector and that $b_k \geq 0$, $k = 1, 2, \ldots, N$, and assume that $x = (x_1, x_2, \ldots, x_N)^*$ is the solution of the equation (6). Then $x_k \geq 0$ for $k = 1, 2, \ldots, N$, $\max_i \{x_i\} \leq \max_i \{b_i\}$ and $\max_i \{x_i\}$ is assumed for an $\underline{i}$ such that $b_i > 0$.

Proof. Equation (6) may be rewritten as

$$\mu (b_i - x_i) + \alpha \sum (x_j - x_i) = 0; \quad i = 1, 2, \ldots, N \quad (\alpha > 0, \ \mu \geq 0) \tag{8}$$

where for each $\underline{i}$ the sum is extended over j's in a subset (different for each $\underline{i}$) of the indices 1, 2, ..., N.

Now assume that x_i is the largest one of the x's. If $x_i > b_i$, then the term $\mu (b_i - x_i)$ is negative and all other terms in (8) are non-positive (as $x_i \geq x_j$ for all j) and so the total sum cannot be equal to zero. Thus, x_i cannot be $> b_i$. For similar reasons, the corresponding b_i cannot be $= 0$, nor can the $\underline{\text{smallest}}$ of the x's be < 0. (Note that the largest of the x's must be > 0, for we have assumed that there exists at least one $b_i \neq 0$.)

It is quite easy to see that the theorem above is valid also for an equilibrium solution of (1).

We now turn to the question how to solve the equation (6) numerically. An approximate solution of this equation can be obtained in many different ways. In general, approximate solutions to linear equations are very difficult to compute, if the number of unknowns is large. Even the fastest computers cannot in a reasonable lapse of time solve linear equations with more than some thousand unknown variables, when the constants in the

equations are arbitrary. In our equation (7), however, we are lucky to have a very special kind of matrix which makes the situation more hopeful. The matrix in (7) is generally extremely "sparce", i.e., almost all elements except a comparatively small number of elements are equal to zero.

In searching for a procedure to solve the equation (6) numerically it is very natural to take the iterative process implied by (3) as a starting point. We have not investigated other procedures, but a small-scale test on the Swedish BESK computer shows that (3) yields quite convenient algorithms.

The next theorem gives necessary and sufficient conditions for the iteration to converge.

Theorem 3. Let a(n) be a vector recursively defined by

$$a(n+1) = a(n) + \Delta t \{\mu b + (\alpha A' - \mu E)a(n)\}; \quad n = 0, 1, \ldots \tag{9}$$

where a(0) is given and Δt, α and μ are positive constants. (A' is the matrix defined in equation (3).) The sequence a(n), n = 1, 2, ... converges when

$$\Delta t \{\mu + 2\alpha m\} < 1, \tag{10}$$

m is defined as the largest number of arrows attached to any single article (irrespective of their direction). The convergence is optimal when

$$\Delta t (\mu + \alpha m) = 1. \tag{11}$$

Proof. Let $a(\infty)$ be the solution of equation (6), (i.e., the steady state solution). Define the vector $a'(n)$ by $a(n) = a(\infty) + a'(n)$. Introducing this into (9) we get the following recursive formula for the vector $a'(n)$

$$a'(n+1) = a'(n) + \Delta t [\alpha A' - \mu E] a'(n).$$

In coordinate form this formula may be rewritten as

$$a_i'(n+1) = [1 - \Delta t (\mu + \alpha m_i)] a_i'(n) + \Delta t \cdot \alpha \sum a_j'(n); \tag{12}$$

$i = 1, 2, 3, \ldots, N$

Again the sum is extended over j´s in a subset of the indices 1, 2, ..., N. m_i is the number of terms in that sum.

Now, if we assume that $[1 - \Delta t(\mu + \alpha m)] \geq 0$, we immediately get from (12) that (note that $m_i \leq m$)

$$|a_i'(n+1)| \leq (1 - \Delta t \cdot \mu) \cdot \max\{|a_i'(n)|\} \tag{13}$$

Otherwise we get the following estimate

$$|a_i'(n+1)| \leq [\Delta t \cdot (2\alpha m + \mu) - 1] \cdot \max\{|a_i'(n)|\}$$

From this last inequality it follows that the mapping[1)] $x \to \{E + \Delta t(\alpha A' - \mu E)\}x$, $\underline{x}$ realvalued vector, is a contraction when the inequality (10) is fulfilled. From (13) it follows that the contraction is as large as possible when (11) is fulfilled. Thus the theorem is proved.

---

1) A mapping $x \to T(x)$ from one Banach space to another is called a contraction if an inequality of the type

$||T(x)|| \leq C \cdot ||x||$, where C is a constant < 1,

is fulfilled. In our case we have a linear transform from R_N to R_N, and we have employed the maximum norm as norm in R_N.